Housing

Brett Sutow

10/14/2020

All questions answered in 1-6 format with output. Due to the length of output for residuals, leverage, cooks distance, and covariance the results were not included in word document. However, the formulas were. Please run in rstudio to see results or view pdf for entire results.

## Set the working directory to the root of your DSC 520 directory and Set dataframe

1. Cleaning Datapoints based off the information and the given datapoints we would clean: City Name, Property Type. These have information that is not full, and not relevant.

## Creating Variables: Sale date picked as another predictor because it allows for comparison based on time

sale\_lm <- lm(housing\_df$`Sale Price` ~ housing\_df$sq\_ft\_lot)  
summary(sale\_lm)

##   
## Call:  
## lm(formula = housing\_df$`Sale Price` ~ housing\_df$sq\_ft\_lot)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2016064 -194842 -63293 91565 3735109   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.418e+05 3.800e+03 168.90 <2e-16 \*\*\*  
## housing\_df$sq\_ft\_lot 8.510e-01 6.217e-02 13.69 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 401500 on 12863 degrees of freedom  
## Multiple R-squared: 0.01435, Adjusted R-squared: 0.01428   
## F-statistic: 187.3 on 1 and 12863 DF, p-value: < 2.2e-16

saleonbed\_lm <- lm(housing\_df$`Sale Price` ~ housing\_df$sq\_ft\_lot + housing\_df$bedrooms)  
summary(saleonbed\_lm)

##   
## Call:  
## lm(formula = housing\_df$`Sale Price` ~ housing\_df$sq\_ft\_lot +   
## housing\_df$bedrooms)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2051983 -171800 -59122 78362 3782453   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.897e+05 1.415e+04 20.48 <2e-16 \*\*\*  
## housing\_df$sq\_ft\_lot 7.772e-01 6.070e-02 12.80 <2e-16 \*\*\*  
## housing\_df$bedrooms 1.017e+05 3.944e+03 25.78 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 391500 on 12862 degrees of freedom  
## Multiple R-squared: 0.06278, Adjusted R-squared: 0.06264   
## F-statistic: 430.8 on 2 and 12862 DF, p-value: < 2.2e-16

## Including Plots

1. Rsquared represents the correlation between the dependent variable and the independent variable. In this case it shows for the sale\_lm an rsquared of .01435 and the rsquared for the model saleonbed\_lm .06278. The addition of the sale year helped improve the rsquared.Showing a better correlation based off that, but rsquare is still very low for us to determine impact between the variables.

## Beta

library(lm.beta)  
lm.beta(sale\_lm)

##   
## Call:  
## lm(formula = housing\_df$`Sale Price` ~ housing\_df$sq\_ft\_lot)  
##   
## Standardized Coefficients::  
## (Intercept) housing\_df$sq\_ft\_lot   
## 0.0000000 0.1198122

lm.beta(saleonbed\_lm)

##   
## Call:  
## lm(formula = housing\_df$`Sale Price` ~ housing\_df$sq\_ft\_lot +   
## housing\_df$bedrooms)  
##   
## Standardized Coefficients::  
## (Intercept) housing\_df$sq\_ft\_lot housing\_df$bedrooms   
## 0.0000000 0.1094197 0.2203058

1. #The beta coefficients in for these datapoints shows a positive relationship.This tells us that as the independent variable increases so does the dependent.These betas are not very strong, as shown in the results. But, they are positive in result.

## Confidence Interval

confint(sale\_lm)

## 2.5 % 97.5 %  
## (Intercept) 6.343730e+05 6.492698e+05  
## housing\_df$sq\_ft\_lot 7.291208e-01 9.728641e-01

confint(saleonbed\_lm)

## 2.5 % 97.5 %  
## (Intercept) 2.620020e+05 3.174787e+05  
## housing\_df$sq\_ft\_lot 6.581999e-01 8.961541e-01  
## housing\_df$bedrooms 9.395180e+04 1.094148e+05

1. We are 97.5% confident that our data points from the created variables will fall within the ranges listed in the charts produced.

## ANova

anova(sale\_lm)

## Analysis of Variance Table  
##   
## Response: housing\_df$`Sale Price`  
## Df Sum Sq Mean Sq F value Pr(>F)   
## housing\_df$sq\_ft\_lot 1 3.0197e+13 3.0197e+13 187.34 < 2.2e-16 \*\*\*  
## Residuals 12863 2.0734e+15 1.6119e+11   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(saleonbed\_lm)

## Analysis of Variance Table  
##   
## Response: housing\_df$`Sale Price`  
## Df Sum Sq Mean Sq F value Pr(>F)   
## housing\_df$sq\_ft\_lot 1 3.0197e+13 3.0197e+13 197.00 < 2.2e-16 \*\*\*  
## housing\_df$bedrooms 1 1.0187e+14 1.0187e+14 664.59 < 2.2e-16 \*\*\*  
## Residuals 12862 1.9715e+15 1.5328e+11   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

5.The residuals decreased from the original variable to the new created one.This means that means there is less of a relationship due to this addition.

## Residuals Output

residsale\_lm<- resid(sale\_lm)  
residsaleonbed\_lm<-resid(saleonbed\_lm)  
rstansale\_lm<-rstandard(sale\_lm)  
rstansaleonyear\_lm<-rstandard(saleonbed\_lm)  
print(rstansale\_lm)  
print(residsaleonbed\_lm)  
sum(residsale\_lm, na.rm = TRUE)  
sum(residsaleonbed\_lm, na.rm= TRUE)

## Leverage

hatvalues(sale\_lm)  
hatvalues(saleonbed\_lm)

## Cooks Distance and Covariance

cooks.distance(sale\_lm)  
cooks.distance(saleonbed\_lm)  
cov(housing\_df$`Sale Price`, housing\_df$sq\_ft\_lot)  
cov(housing\_df$`Sale Price`,housing\_df$bedrooms)

## Indpendence and Multiccollinearity

function(sale\_lm)  
function(saleonbed\_lm)  
library(MASS)

## function(sale\_lm)  
## function(saleonbed\_lm)  
## library(MASS)

tbl= table(housing\_df$`Sale Price`, housing\_df$sq\_ft\_lot)  
tbl1= table(housing\_df$`Sale Price`, housing\_df$bedrooms)  
chisq.test(tbl)

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: tbl  
## X-squared = 24845528, df = 24256666, p-value < 2.2e-16

chisq.test(tbl1)

## Warning in chisq.test(tbl1): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: tbl1  
## X-squared = 54950, df = 44198, p-value < 2.2e-16

library(car)

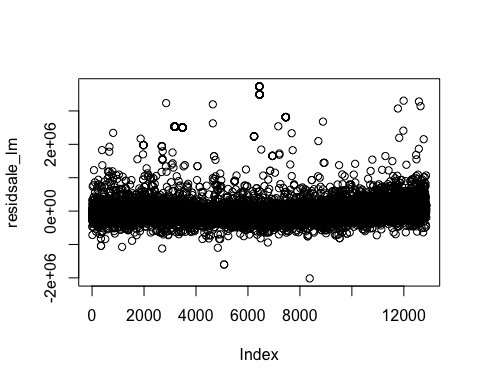
## Loading required package: carData

vif(saleonbed\_lm)

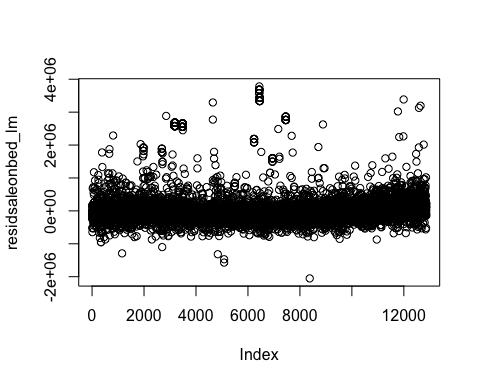
## housing\_df$sq\_ft\_lot housing\_df$bedrooms   
## 1.00223 1.00223

## Plot

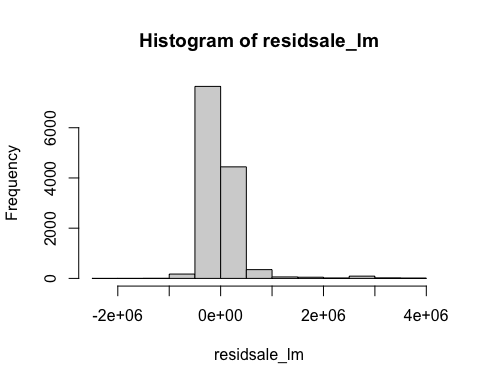
plot(residsale\_lm)



plot(residsaleonbed\_lm)



hist(residsale\_lm)



hist(residsaleonbed\_lm)



6. Based off the information above I believe the data is not biased. Though one of the key factors we can look at is multicollinearity and it is above 1.It is not above it substainally enough to show significant impact on the data.